Fuzzy Logic Inference System for Voiced-Unvoiced-Silence Classification of Malaysian English Isolated Words based on Time-domain Features

Yusnita MA[1] *Paulraj MP*[2] *Sazali Yaacob*[2] *Shahriman AB*[2] *Satheesh Kumar Nataraj*[2]

Abstract

The ability of a system to automatically detect speech signals such as, voiced-unvoiced-silence can have a great impact to the accuracy and computation time in isolated-word speech recognition systems. The presence of a long inter-syllable silence in Malaysian English speech is due to the habit of spelling and reading methods in syllabic manner of the national Malay language at school. Further, manual segmentation is very tedious, laborious and error-prone for large vocabularies and sample size. Time-domain features such as short-time energy and zero-crossing rate are proven to be very efficient and simple computations. Since frame-based speech analysis does not consider the start and end of a particular phoneme, the best and easiest way to make decision for the important features should be inferred from human expert knowledge. In this paper, fuzzy inference system is build from membership functions and fuzzy rules derived from a simple statistical analysis of the speech data from males and females and three ethnic groups of Malaysian English speakers. The study of voiced-unvoiced-silence classification using fuzzy logic method with energy and zero-crossing rate input vectors are based on the simplest settings of three two-syllabic words and under noise-free environment. The experimental results show highest accuracy rates of 99% for voiced speech and 94% for silence.

Keywords – Voiced/Unvoiced/Silence classification, Malaysian English, Short-time energy, Zero-crossing rate, Fuzzy logic.

I. Introduction

There are basically two methods to segment speech signals. Traditional automatic speech recognition systems (ASRs) try to separate blocks into the same characteristics which represent a single phoneme, which is the smallest unit of speech construction in a particular language. Malcangi [1] and Tolba [2] performed analysis based on phoneme segmentation. The other method and perhaps most of the modern ASRs practice is using fixed-frame segmentation due to its simplicity and straightforward process. While the earlier is claimed to have better representation, however, it is complicated by the fact that phonemes can blend together under certain circumstances and variety of imperfect articulation of a spoken word could make this task disputable to detect correct boundaries. According to [3, 4], frames are made in the size of 10 ms to 30 ms due to the theory that temporal variation of the vocal tract shape is relatively slow thus maintain a constant characteristics within this time frame length. This is also a compulsory condition for using fast Fourier transform and Linear Predictive Coding (LPC) techniques. In speech preprocessing, automatic speech segmentation is a key processing task to overcome manual

[1]Faculty of Electrical Engineering, Universiti Teknologi MARA Malaysia, Shah Alam, Malaysia. Email: yusnita082@ppinang.uitm.edu.my

[2]School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh, Malaysia. Email: paul@unimap.edu.my

segmentation which is time consuming, laborious, tedious, error prone and expert dependent. The classification of speech into voiced, unvoiced and silence abbreviated as V-UV-SIL or the decision of endpoint detection has been intensely researched in the field of speech applications such as speech recognition, synthesis and coding. Not all key information embedded in the pre-recorded speech is contained in the short-time frames of the uttered speech. Discarding the unnecessary frames of non-speech part is preliminarily important to have an efficiently less computational and more accurate isolated-word ASRs. These systems are used primarily for handling voice commands such as in voice dialing and security control to access confidential information area.

The model of speech can be best described using source-filter model [5] which describes speech sounds as the product of convolution between the vocal tract filter and the glottal source and assuming that these components are independent from each other, giving it the second name as linear separable model as well. The mechanism of having voiced speech is the vibration of the vocal folds in response to airflow from the lungs and it is periodic in nature. Conversely unvoiced speech is caused by turbulent airflow due to a constriction in the vocal tract and exhibited by a noise-like signal. All vowels and semivowels sounds and some consonants are voiced while most of the consonants are unvoiced. The vowels are 12 dB louder than consonants [6] in general. Malaysian English (MalE) is colored by different pronunciations as it is influenced by various ethnicities [7], thus complicating the structure in comparison to native English pronunciations such as British English. Some voiced sounds like /z/ is pronounced unvoiced and a Chinese Malaysian replaces the phoneme with voiced /£/. Final unvoiced plosive /t/ is realized as unvoiced glottal stop /"/ for a Malay speaker. MalE speakers tend to speak in a syllabic manner (syllable-stress) as they are taught to spell and read by that manner. As the result they tend to prolong the gap between syllables in a word.

The advent of fast computer hardware and algorithms has witnessed many techniques implemented to achieve this goal. The appropriate ways of extracting features of a spoken word such as simple time-domain features namely short-time energy (STE), zero-crossing rate (ZCR) and pitch contour. However, pitch has some drawback [8] as it depends very much on voice periodicity wherein a single frame might be just quasi-periodic or non-periodic. Pattern recognition technique and statistical decision theory have been successfully used such as in [8, 9] in determining the decision threshold. In [9] three algorithms were developed based on histogram of STE only for speech detection. Multiple features such as log STE, ZCR, first autocorrelation coefficient, first LPC coefficient and log energy of prediction error were used in [8] to calculate mean vector and covariance matrix of probability density functions of each V-UV-SIL class. Based on lower and upper threshold limits of energy of speech and ZCR threshold of silence, a backward and forward search algorithm for endpoint detection was establish in a well known paper [10] for isolated utterances. This algorithm was further investigated [11] and features were extended using frame-based Teager's energy and energy-entropy to compare the performance in detecting Malay isolated-words speech recognition. An even simple decision barely made based on just the amount of energy and ZCR count algorithm was establish in [12] for the word '*four*'. Spectral features are less encountered in V-UV decision probably due to its more computation cost but it was reported in [13] using Mel frequency cepstral coefficient and LPC with Gaussian mixture model. The attempt to apply fuzzy logic to V-UV-SIL is quite new although it has been applied to segment phonetic unit of vowels and consonants in [1] and voiced-silence classification for pathological speech in [14]. Thus separating V-UV from SIL using simple features namely log STE and ZCR and fuzzy logic as simple and human-like rules predictor is inevitably important in speech processing of MalE speech which is the focus of the study in this paper. The attempt to evaluate the developed system in comparison to manual labeling is also presented.

This paper is organized as follows. In section II, a brief description about MalE speech database is presented. Section III describes methodologies used for extracting speech features and classifier used to do the classification of V-UV-SIL frames. The experimental setup and findings are discussed in section IV. Lastly, section V gives the concluding remarks based on the studied problem and methodology.

II. Malaysian English Speech Database

For the purpose of analysis, speech utterances recorded from seven males and females from two major ethnicities i.e. Malay and Chinese and from seven males and five females Indian MalE speakers uttering three two-syllable words i.e. *bottom*, *student* and *zero*. Two-syllabic words were selected to prove that some words have great inter-silence. Each word was replicated five times to generate more samples from a speaker. The average duration of each word is about 0.5 s to 0.7 s and was hand-segmented prior to analysis. Subjects were postgraduate students of Universiti Malaysia Perlis aged from 18 to less than 30 years. The dataset contains total utterances of 200 samples for each word under the study. The recording was carried in an acoustic chamber room which is semi-anechoic using a handheld condenser and unidirectional microphone. The background noise in that room was approximately 22 dB. This level is considered very quiet and controlled as compared to normal quiet room about 40 to 50 dB. The speech was recorded using a laptop computer sound card and MATLAB program where the sampling rate was set to 16 kHz and bit resolution was set to 16 bps.

III. Methodology

Firstly, the DC components captured in the raw data due to microphone setup was removed or zero-adjusted. Next, frame-blocking into 256-point frame with 128-point overlapping between consecutive frames were made on each word sample prior to log STE and ZCR extractions. The flowchart in Fig. 1 summarizes the process of classifying V-UV-SIL speech

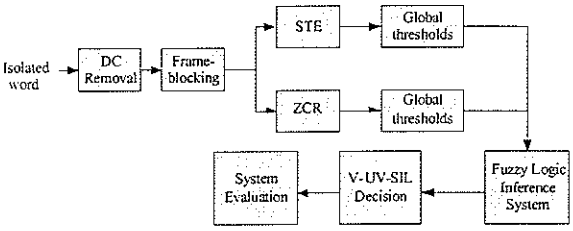

Figure 1. Block Diagram of V-UV-SIL System

A. Short-time Energy

Time-domain feature such as STE is efficient and easy to compute, thus simplify the hardware implementation. Higher energy is an indicator of voiced speech as unvoiced is spoken at less power and very much less energy should be in silence. Energy is associated positively with loudness of the acoustic signal and also related negatively to its frequency. It is found that the spectrum decreases in amplitude with increasing frequency at a rate of around 12dB per octave. This feature measures the sum of squared magnitude of the sample values in each frame and can be expressed as in (1) as log-energy.

$$E(k,n) = 10 * \log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1}\left|x(k,n)\right|^{2}\right) \qquad (1)$$

where variables *E(.)* and *x(.)* represent frame-energy and input speech respectively and *k* is the k^{th} frame and *n* is the n^{th} sample point in a frame. *N* is the length of the frame.

B. Zero-crossing Rate

Another simple yet powerful time-domain feature is ZCR. It is a measure of how many times the waveform crosses the zero-axis in a particular frame. Also, it can be a crude estimation of pitch determination wherein the number of crossings per second is equal to twice the frequency and best used in the absence of noise environment. Generally, the ZCR of both unvoiced speech and background noise are higher than voiced speech which has obvious fundamental periods and has most of the energy concentrated at low frequencies. The equation to count zero-crossings is given in (2).

$$ZCR(k,n) = \frac{1}{2N}\sum_{n=0}^{N-1}\left|sign\left[x(k,n)\right] - sign\left[x(k,n-1)\right]\right| \quad (2)$$

where variables $ZCR(.)$ and $x(.)$ represent frame-ZCR and input speech respectively and k is the kth frame and n is the nth sample point in a frame. N is the length of the frame. The function $sign[.]$ is defined as follow

$$sign\left[x(n)\right] = \begin{cases} +1 \; if \;\; x(n) \geq 0 \\ -1 \; if \;\; x(n) < 0 \end{cases} \qquad (3)$$

C. Fuzzy Logic Inference System

Fuzzy logic was first introduced by Zadeh in 1965 who defined fuzzy set theory to describe fuzziness. In Zadeh's second most influential paper [15], human knowledge was suggested to be captured in fuzzy rules. Thus, fuzzy logic leads to more human-like intelligent system as it models our sense of word and decision making based on common sense. It is basically a set of mathematical principles for knowledge representation based on degrees of membership [16]. The behavior of the collected speech data from MalE utterances was studied using statistical analysis (minimum, mean and maximum values) of STE and ZCR measures to derive fuzzy membership functions and fuzzy rules. To ensure that the rules fit our data, the inferred knowledge was also based on experimental observations that correlate to these three classes to tune the fuzzy engine. This system was designed using MATLAB fuzzy logic toolbox GUIs. Manual hand-labeling of V-UV-SIL of frames of each word were made to evaluate the fuzzy inference system (FIS). Since this task was laborious and time-consuming, only three words were analyzed. Fig. 2(a)–(c) depict the membership functions for STE and ZCR inputs and speech output for the word '*bottom*'. A triangular membership function was chosen as it is practical and one of the simplest linear-fit functions. To keep the design simple, a minimum number of three membership functions in input and output were generated.

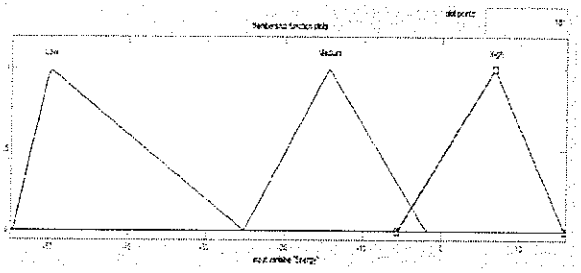

Figure 2(a). Membership function for STE

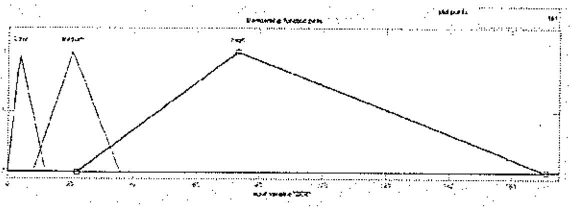

Figure 2(b). Membership function for ZCR

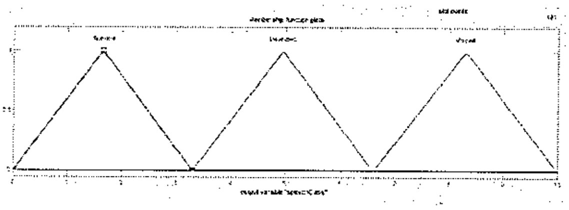

Figure 2(c). Membership function for system output

There were five steps involved in developing FIS. (1) Fuzzification of inputs–the crisp measures STE and ZCR were taken to determine their degree of membership in each fuzzy subset. (2) Application of fuzzy operators–only AND operators were used in five rules using *min* built-in method. (3) Application of implication method–*min* was used to truncate output fuzzy sets. (4) Aggregation of all outputs–fuzzy sets that represents the outputs of each rules were combined using *max* method into a single fuzzy set. Lastly (5) Defuzzification–resolves the single fuzzy set to a single crisp value.

The fuzzy rules in V-UV-SIL FIS are tabulated in Table 1.

TABLE 1. A SET OF RULES for V-UV-S SYSTEM

Rule#		STE	Operator	ZCR	Output
1	if	LOW	and	NONE	SIL
2	if	MEDIUM	and	MEDIUM	V
3	if	MEDIUM	and	HIGH	UV
4	if	MEDIUM	and	LOW	V
5	if	HIGH	and	NONE	V

IV. Results & Discussion

In this section, the experimental setup and results are reported. Three experiments based on three isolated words i.e. '*bottom*', '*student*' and '*zero*' were conducted to find the thresholds for generating fuzzy membership functions. Each word consists of the following phonemes sequences according to British English:

bottom - /'b, , t, ә, m/

student - /'s, t, j, u:, d, ә n, t/

zero - /'z, Iә, r, ә /

Only a great detail of the word '*bottom*' is shown and discussed as example for analyzing the results. Fig. 3 depicts the waveform '*bottom*' consisting of its phonemes sequence and silence.

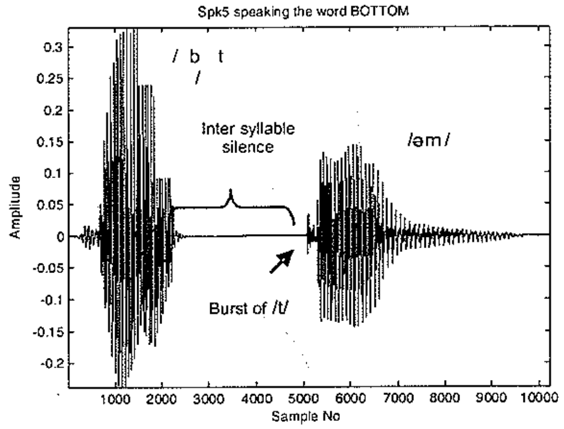

Figure 3. The waveform of Spk 5 speaking the word 'bottom'

It is obvious that the MalE speech is characterized by syllabic stress and a long inter-syllable silence which can cause a waste in processing. The silence was produced before the release of plosive /t/. Table 2 and 3 tabulate the results of STE and ZCR measures of ten test samples of '*bottom*'. This test dataset is 5% fraction of the available data.

TABLE 2. TEST DATASET WITH STE MEASURES

No	Speaker ID	Gender	Ethnic	Min STE	Mean STE	Max STE
1	Spk5	Female	Malay	-50.11	-20.18	5.67
2	Spk7	Male	Indian	-52.23	-15.01	9.59
3	Spk13	Male	Malay	-49.93	-8.85	6.83
4	Spk19	Male	Chinese	-50.59	-11.12	10.27
5	Spk25	Female	Chinese	-53.18	-20.31	2.81
6	Spk95	Female	Indian	-48.70	-14.04	8.5
7	Spk112	Male	Chinese	-51.88	-14.29	11.04
8	Spk141	Female	Indian	-53.12	-11.93	6.00
9	Spk97	Female	Malay	-50.74	-10.84	10.70
10	Spk81	Male	Malay	-51.74	-12.65	8.73

TABLE 3. TEST DATASET WITH ZCR MEASURES

No	Speaker ID	Gender	Ethnic	Min STE	Mean STE	Max STE
1	Spk5	Female	Malay	0	17.56	59
2	Spk7	Male	Indian	3	23.71	97
3	Spk13	Male	Malay	0	14.38	73
4	Spk19	Male	Chinese	3	16.58	60
5	Spk25	Female	Chinese	7	30.51	83
6	Spk95	Female	Indian	0	15.91	51
7	Spk112	Male	Chinese	2	17.23	92
8	Spk141	Female	Indian	6	23.05	80
9	Spk97	Female	Malay	0	14.35	64
10	Spk81	Male	Malay	6	22.55	86

Next, the frames having different ranges of STE and ZCR values were analyzed to discuss the results for Spk5. Fuzzy rules generated in Table 1 were justified based on manual observations and the computed features values in Figure 4(a)–(f).

It was observed that the waveform in Fig. 4(a) i.e. frame 28 displayed as noise-like signal, was the inter-syllable silence between the first syllable /'bRt/ and the release of /t/ at its final position in the speech. These characteristics justified Rule 1 in the FIS.

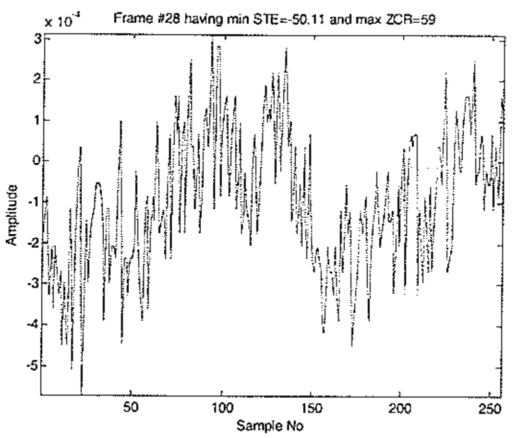

Figure 4(a). Frame #28 of Spk5 with min STE and max ZCR

Fig. 4(b) is another inter-syllable silence but with very low STE and minimum silence (ZCR = 0). This condition might be due to the microphone setting. These characteristics also justified Rule 1 in the FIS.

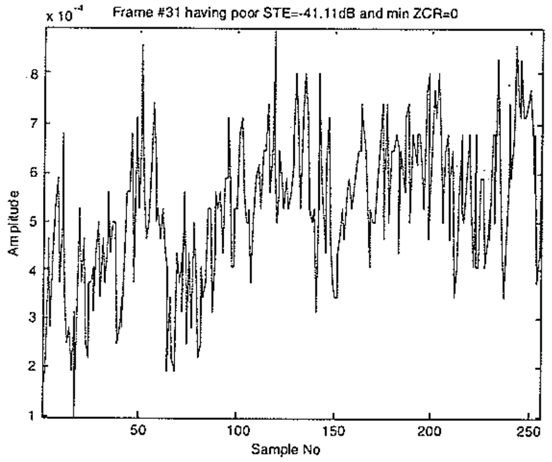

Figure 4(b). Frame #31 of Spk5 with low STE and min ZCR

Frame 9 in Fig 4(c) justified Rule 5 with max STE and medium ZCR. All frames with high STE regardless of ZCR count was considered voiced part. This frame was a voiced open back rounded short vowel /R/ in the word as characterized by the stable pitch of multiple combinations of frequencies.

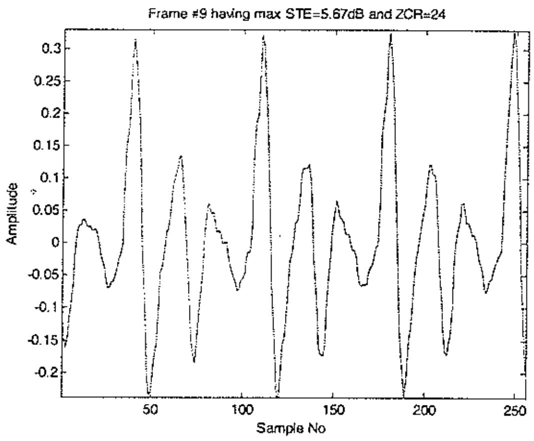

Figure 4(c). Frame #9 of Spk5 with max STE and medium ZCR

Frame 71 in Fig. 4(d) originated from phoneme /m/ and it was a voiced nasal consonant characterized by a single frequency sound. These characteristics justified Rule 4 in the FIS.

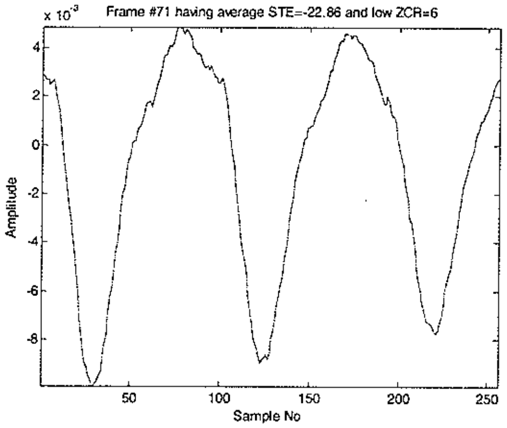

Figure 4(d). Frame #71 of Spk5 with medium STE and low ZCR

Frame 63 in Fig. 4(e) originated from voiced vowel schwa /Y/. These characteristics justified Rule 2 in the FIS.

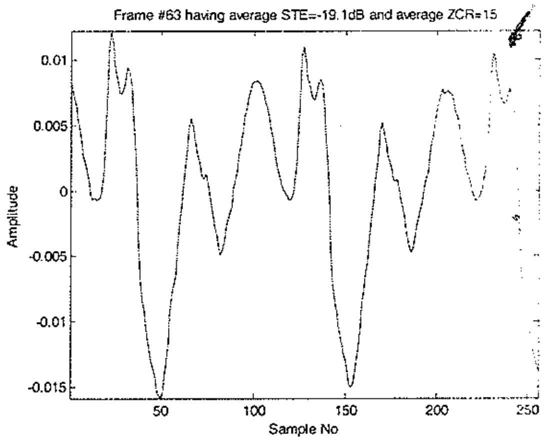

Figure 4(e). Frame #63 of Spk5 with medium STE and medium ZCR

Lastly the analysis for fuzzy rule involved Rule 3, wherein frame 40 in Fig. 4(f) was postulated to originate from a rapid burst of unvoiced consonant /t/ preceding the voiced vowel schwa /Y/.

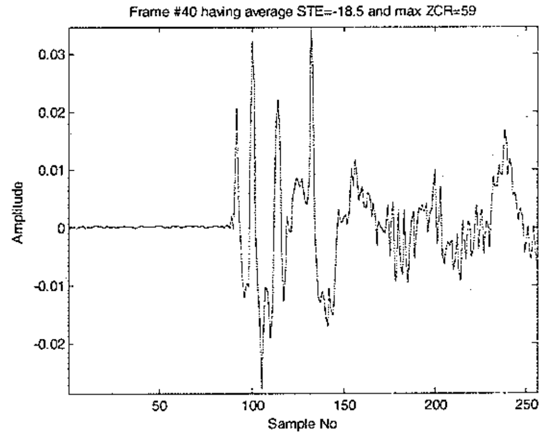

Figure 4(f). Frame #40 of Spk5 with medium STE and max ZCR

For the purpose of system evaluation, accuracy rate was computed based on the difference between manual labeling of human expert and FIS output. The results are tabulated in Fig. 5. With reference to the experimental results, the overall classification rates for bottom, student and zero are 92.64%, 92.22% and 92.27% respectively. The breakdown of accuracy rates for voiced, unvoiced and silence is presented as bar chart and clustered across each word.

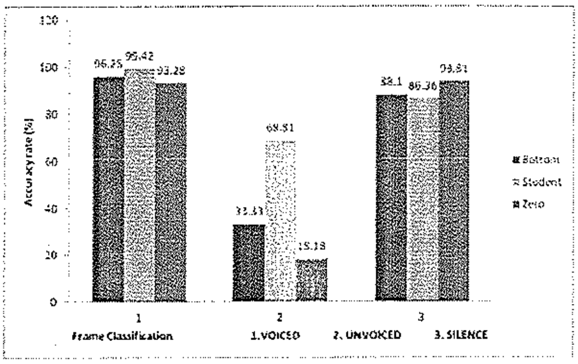

Figure 5. Accuracy rate for classification of V-UV-S for three isolated words

In further analysis, the accuracy rates for speech and non-speech were obtained by combining the voiced and unvoiced frames. By doing this, the averagely low accuracy rate of unvoiced can be solved. It is proven that frame reduction is achieved by eliminating the inter-syllable silence. The results for speech and non-speech classification and frame reduction rates of each word can be visualized in Fig. 6 and Fig. 7.

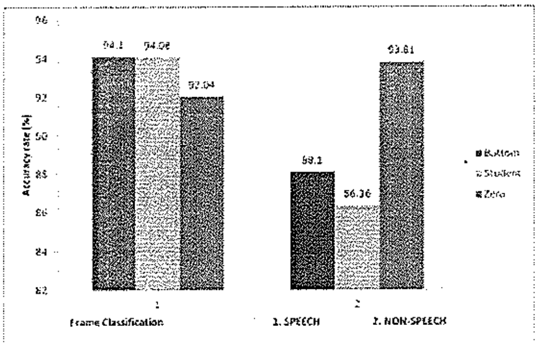

Figure 6. Accuracy rate for classification of speech and non-speech for three isolated words

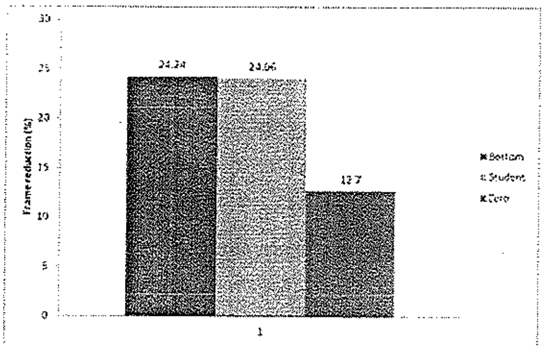

Figure 7. Frame reduction rate for three isolated words

V. Conclusion and Future Work

This paper has presented a classification of speech into V-UV-SIL of MalE isolated words using an intelligent FIS. Based on the distribution of crisp data from short-time log-energy and zero-crossing rate and using simple statistical data analysis, membership functions and fuzzy rules were derived to build the system. Classification results shows that FIS has demonstrated a great potential in detecting voiced speech with a high accuracy rate ranging from 93% ('*zero*') to 99% ('*bottom*'), followed by silence 88% ('*bottom*') to 94% ('*zero*'). However, it is a non-trivial problem to detect unvoiced from voiced and silence because of its partially low energy and overlap characteristic of zero-crossings like silence. The accuracy obtained was in between 18% ('*zero*') to 67% ('*student*'). However the accuracy rate of detecting speech and non-speech parts are satisfactory i.e. in average 93.4% and 89.4% respectively. This should enable to capture the important speech features and at the same time, resulted in speech compression of as much as 24% via eliminating the inter-syllable silence problem which has been shown quite significant in MalE speech. The plan for future work is to demonstrate that classification of accent types of MalE speakers can be more efficient using the results of this V-UV-SIL classification.

Acknowledgment

The authors would like to acknowledge the support and encouragement given by the Vice Chancellor of University of Malaysia Perlis (UniMAP), Brig.Jeneral Dato' Prof. Dr.Kamaruddin Hussin.

References

[1] M. Malcangi, "Softcomputing approach to segmentation of speech in phonetic units," *International Journal of Computer and Communications*, vol. 3, pp. 41-48, 2009.

[2] M. F. Tolba, T. Nazmy, A. A. Abdelhamid, and M. E. Gadallah, "A novel method for Arabic consonant/vowel segmentation using wavelet transform," *International Journal on Intelligent Cooperative Information Systems, IJICIS*, vol. 5, pp. 353-364, 2005.

[3] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, vol. 103: Prentice hall Englewood Cliffs, New Jersey, 1993.

[4] S. Furui, *Digital speech processing, synthesis, and recognition*, vol. 7: CRC, 2001.

[5] S. Cassidy. (2000). Chapter 7. The Source Filter Model of Speech Production. [Online]. Available: http://web.science.mq.edu.au/~cassidy/comp449/html/ch07.html.

[6] I. McLoughlin, "Applied Speech and Audio Processing," New York: Cambridge University Press, 2009.

[7] S. Nair Venugopal, "English, identity and the Malaysian workplace," *World Englishes*, vol. 19, pp. 205-213, 2000.

[8] B. S. Atal and L. R. Rabiner, "Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, pp. 201-212, 1976.

[9] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Fifth European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[10] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, pp. 297-315, 1975.

[11] N. Seman, Z. A. Bakar, and N. A. Bakar, "An evaluation of endpoint detection measures for malay speech recognition of an isolated words," in *Information Technology (ITSim), 2010 International Symposium in*, 2010, vol. 3, pp. 1628-1635.

[12] K. Elleithy, R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," in *Advanced Techniques in Computing Sciences and Software Engineering*: Springer Netherlands, pp. 279-282.

[13] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Robust voiced/unvoiced classification using novel features and gaussian mixture model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 17-21.

[14] M. P. Paulraj, S. Bin Yaacob, A. N. Abdullah, and S. K. Natraj, "Fuzzy voice segment classifier for voice pathology classification," in *Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium on*, 2010, pp. 190-195.

[15] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, pp. 28-44, 1973.

[16] M. Negnevitsky, *Artificial intelligence: a guide to intelligent systems*, Third ed. Edinburgh: Pearson Education Limited, 2011.

Author's Biography

Yusnita Mohd Ali received her Master Degree in Electronics System Design Engineering from University Sains Malaysia in 2004. She completed her Bachelor Degree. In Electrical & Electronics Engineering from the same university in 1998. Currently, she is pursuing Ph.D. study in Universiti Malaysia Perlis and at the same time she is a lecturer in Universiti Teknologi MARA Malaysia. Her field of interest is speech and accent recognition, signal processing and artificial intelligence.

Dr. Paulraj MP received his BE in Electrical and Electronics Engineering from Madras University (1983), Master of Engineering in Computer Science and Engineering (1991) as well as Ph.D. in Computer Science from Bharathiyar University (2001), India. He is currently working as an Associate Professor in the School of Mechatronic Engineering, University Malaysia Perlis, Malaysia. His research interests include Principle, Analysis and Design of Intelligent Learning Algorithms, Brain Machine Interfacing, Dynamic Human Movement Analysis, Fuzzy Systems, and Acoustic Applications. He has co-authored a book on neural networks and 250 contributions in international journals and conference papers. He is a member of IEEE, member of the Institute of Engineers (India) and a life member in the System Society of India.

Dr. Sazali Yaacob received his Ph.D. Degree in Automatic Control and System Engineering from University of Sheffield, UK in 1995. He is working as a Professor & appointed as Head of Intelligence Signal Processing Group in Universiti Malaysia Perlis. He is interested in human behavior modeling, automatic control system and smart satellite system. He has published more than 180 international/national conference papers and more than 66 journal papers, 7 book chapters and 4 academic books